PRE-PROCESSING OF INPUT FEATURES USING LPC AND WARPING PROCESS

Rubita Sudirman, Sh-Hussain Salleh, Ting Chee Ming
Biomedical Engineering Research Group
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
email: rubita@fke.utm.my

ABSTRACT

This paper presents pre-processing of input features to artificial neural network (NN). This is for preparation of reliable reference templates for the set of words to be recognized. The first task is to extract pitch features using Pitch Scale Harmonic Filter (PSHF) algorithm. Another task is to align the input frames (test set) to the reference template (training set) using a modified DTW algorithm called DTW fixing frame (DTW-FF) algorithm. This proper time normalization is needed since NN is designed to compare data of the same length; same speech can varies in their duration. By performing frame fixing or time normalization, the test set and the training set is adjusted to a fix number of frames throughout the sets utilizing the local distance score of the matched features. Then those features can be adapted to NN for further recognition tuning.

I. INTRODUCTION

Due to time variant, time alignment method has to be employed if NN is installed as the back end speech recognition engine. There are few time alignment methods which is also known as time normalization methods being used by people working in this fields; namely trace segmentation and DTW [1, 5, 6, 8]. They are used to interpolate input signal into a fixed size of input vector. So, input feature processing is going to utilize one of the methods before the feature can be presented into NN system. Many had used either HMM or DTW method with NN [1, 6, 8].

DTW has been one of the prime speech recognition methods since its birth more than 30 years ago, it works by matching the unknown speech input template to a pre-define reference template [4, 12], and this method is an easiest speech recognition method compared to others like HMM or TDNN. TDNN approach dealing with the time alignment problem by mapping speech temporal variation into interconnections of neurons of different delays [8, 10]. However, in this paper, DTW is used as an input feature preprocessor to perform time normalization because of its ability to search the best path between two time-series signals [3]: data is expanded or compressed according to reference template. The proposed pre-processing also applies trace segmentation method to reduce data size, in which the initial idea of trace segmentation is to reduce the number of stored feature vectors for the stationary portion [1, 6]. Our methods can even reduce the feature vector by 90% through the DTW frame alignment method. This is utilizing the local distance score collected from the matching process between input and the reference template. Preliminary test on the aligned frame data using typical DTW shows similar result as before normalization process.

The following sections are brief of feature extraction method, followed by DTW time normalization algorithm which is the main aspect of this paper and finally experimental results and some discussion. Some results of the experiment are presented in the conclusion section.

II. PRE-PROCESSING OF SPEECH SIGNALS

Two input features is considered in this preprocessing stage; pitch (an acoustical property) and LPC coefficients (spectral property). There are many feature extraction methods for speech signals like MFCC, LPC, and LPCC. However, in this work linear predictive (LP) is chosen over other methods due to its ability to encode speech at low bit rate and can provide an accurate speech parameters, so that least information is lost during feature extraction process, and its fast computing speed. It has been widely used by speech researchers as speech features representation, in which its basic concept is that it can use first few samples to represent all speech samples. The linear predictive cepstral derived of 10 coefficients in each 10ms frame window is proposed.

Dynamic Time Warping (DTW)

Template matching is an alternative to perform speech recognition; the template matching encountered problems due to speaking rate variability, in which there exist timing differences between the similar utterances. This matching is similar to time-normalization, in this research it is to time normalize a pattern to a standard duration with respect to other pattern of same word based on a template which has average frame amongst a particular word set.

DTW was first introduced by [3], was used for recognition of isolated words in association with Dynamic Programming (DP). It is a technique to normalize the duration variability. Thus, the problem of time differences can be solved through DTW algorithm, which is by warping the reference template against the test utterance based on their features similarities. So, DTW algorithm actually is a procedure, which combines both warping and distance measurement, based on their local and global distance [8, 12].

A few restrictions have to be applied to the warping function to ensure close approximation of properties of actual time axis variations. This is to

 1-4244-0011-2/05/$20.00 ©2005 IEEE.

preserve essential features of the speech pattern. The warping function slope is more rigidly restricted by increasing slope, but if it is too severe then time normalization is not effective, so a denominator to time normalized distance, N is introduced. However N is independent of the warping function, so the time normalized distant is:

$$D(A,B) = \frac{1}{N} \underset{F}{Min} \left[\frac{\sum_{i=1}^{I} d(i, j(i)) * w(i)}{\sum_{i=1}^{I} w(i)} \right] \text{ and } N = \sum_{i=1}^{I} w(i) \quad (1)$$

Having this time normalized distant, minimization can be achieved by DP principles.

In this research, the time normalization is done based on DTW method by warping the input vectors to reference vector which has an almost similar local distance, while expanding vectors of an input to reference vectors which has vertical movement; shares same feature vectors for a feature vector frame of an unknown input. This frame alignment or also called as time normalization technique is also known as the expansion and compression method [7, 8, 11]. Trace segmentation is not chosen as the normalization technique because of its bad past performance in speech recognition, cannot even provide the same performance as DTW. This is due to inappropriate distance segmentation and spatial sampling rate along the trace [13], in addition to that it can only perform frame reduction during the stationary speech portion. DTW is a nonlinear time normalization (LTN) technique that can perform both frame expansion and reduction, and still can preserve important features during the process. It has been proven that LTN is simple procedure to align variable speech length, only that proper LTN values need to be identified so speech information can be sustained [1]. Thus, DTW is used as the new algorithm base with LPC features of 10 coefficients.

III. DTW-FF ALGORITHM

The DTW-FF algorithm is done by compression and expansion technique. In general, the frame compression ($\boldsymbol{F}^-$) is done when several frames of unknown input match to a single frame of reference template. On the other hand, expansion ($\boldsymbol{F}^+$) is done when a single unknown input frame is matched with few frames of the reference. The frame fixing is done following the slope conditions as described follows. There are three slope conditions that have to be dealt with in this research work based on DTW type 1, w(i) is the input frame vector.

(i) ***Slope is 0 (horizontal movement)***
The frames of speech signal are compressed: This is done by taking the minimum local distance amongst the distance set, i.e.: compare w(i) with w(i-1) and choose the frame with minimum local distance.

(ii) ***Slope is ∞ (vertical movement)***
The frame of speech signal is expanded, i.e.: the reference frame gets the identical frame as w(i) of unknown input.

(iii) ***Slope is 1 (diagonal movement)***
The frame is left as it is because it has the least local distance compared to other movements.

This $\boldsymbol{F}^-$ and $\boldsymbol{F}^+$ is done by using new DTW frame fixing algorithm (DTW-FF). Consider the frame vectors of LPC coefficients for input as i…I, and reference as j….J, while F denotes the frame. Frame compression involves searching minimum local distance out of distances in a frame set within a threshold value, it is represented as

$$F^- = \mathrm{F}(\min\{\mathrm{d}_{(\mathrm{i,j})\ldots(\mathrm{I,J})}\})$$

Frame expansion involves duplicating a particular input frame to multiple reference frames of w(i), represented as

$$F^+ = \mathrm{F}(\mathrm{w(i)})$$

If the uppermost warping path coordinate of a reference pattern was (M,N), then the fixed frame number is equal to M. Having the unknown input being fixed to same frame number as the reference, the fixed-frame local distances are retained. Then these data are ready to be used for neural network recognition.

Doing the process throughout a speech sample will yield to frames of unknown input that being fixed to same number of frames as the reference template. The normalized data/sample has being tested and compared to typical DTW algorithm and results showed a same global distance score. However the local distances are different depending on the normalization that has being imposed. Further findings are discussed in the results and discussion section.

Our small vocabulary Malay speech database is used which consists of utterance of digits 0-9 recorded in 5 sessions uttered 5 times each session by 5 subjects. Recognition using combination of time normalization and NN has been done to same utterances of digits 0-9 in English [2], even in Thai language [7], yield a very high accuracy but for small vocabulary. But [9] used NN as recognition tool with phonetic-based DTW, a smaller unit of a word because it is easier to tackle time variation problem of a smaller unit like a phone, and later they carried out experiment on 100 words and obtained 98% recognition result.

IV. RESULTS AND DISCUSSION

The tokens used are digits 0-9 recorded in Malay language for 5 sessions, uttered five times each session.

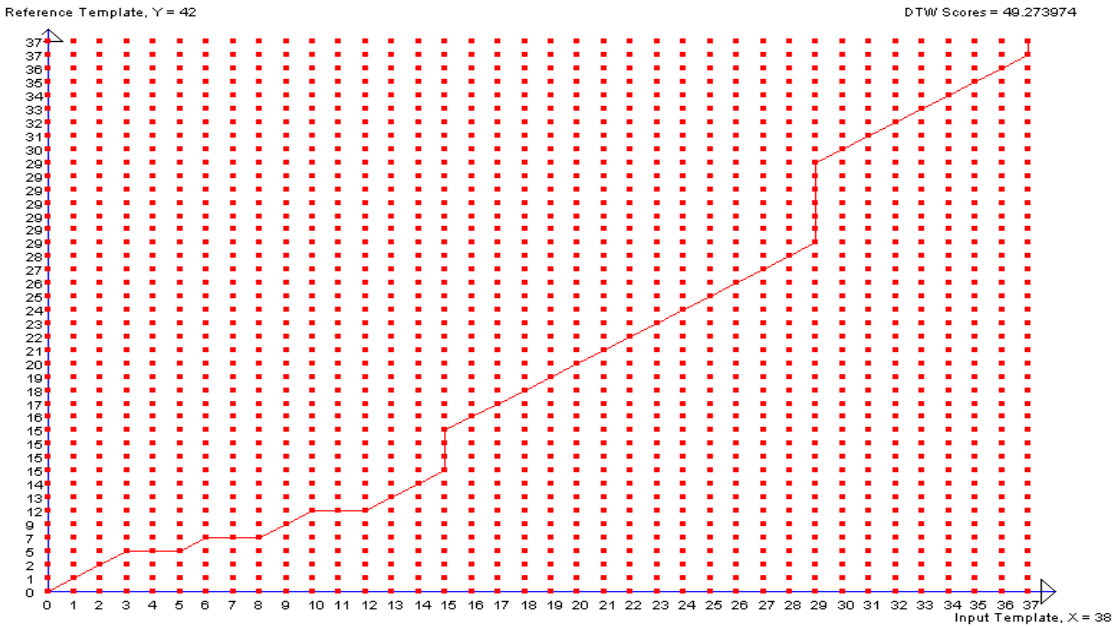


Figure 1: The DTW frame alignment between an input and a reference template of /zero/; the frame number 0-37 (total of 38 frames of input) on reference template axis actually contains 42 frames of reference template itself.

The recognition accuracy can be increased by increasing number of subjects; increase size of database. Figure 1 shows an input frame of digit '0' has been matched to a reference template of same utterance. In this example, initially the input template has 38 frames and reference template 42 frames. Then by using the DTW-FF algorithm the input frames have been expanded to 42, i.e. equals to number of reference frame. According to slope condition (i), considering w(i) as the unknown input frames and r(i) is the reference template frame, local distances of unknown input frame w(3),…,w(5) are compared and w(5) appears to have the minimum local distance among the three frames, so those 3 frames is compressed to one occupies only r(4). On the other hand, slope condition (ii) shows an expansion, for example while w(15) of input are expanded to 4 frames, in which these 4 consecutive frames of the reference template are identical; 4 frames of reference template at r(10),..,r(13) have the same feature vectors as frame w(15) of the input vectors, so w(15) occupies r(10),…,r(13). These means that frame w(15) of the input has matched 4 feature vectors consecutively of the reference set. Since diagonal movement of slope 1 is the fastest track to achieve the global distance and it gives the least local distance at all time compared to the horizontal or vertical movements, a normal DTW procedure is applied to it. Having done expansion and compression along the matching path, the unknown input frame is already matched to reference template frames. In addition, local distance scores are collected for further use to NN, this means instead of feeding the whole frames with 10 coefficients each frame for let say 42 frames which total up to 420 coefficients, the input to NN has been reduced to only the number of local distances which is also equal to the number of the fixed frames, which is 42. Our next step will seek the use of reduced inputs presented to NN, means reducing network complexity, computation time and storage capacity. Early NN experiment using the local distance score shows a good improvement to the recognition up to 100% using speaker dependent samples of 100 utterances.

Further, pitch information will be presented to NN as another speech feature under investigation. Fixed frame speech pattern can also be another character to further enhance the speech recognition besides their local distance scores and any other coefficients that might be used, especially when there exist similarities between two utterance, for example 'tiga' as for digit '3' and 'tujuh' as for digit '7' in Malay language, look at Figure 2 and observe the similarities at the beginning until almost three quarter of the utterance and its little differences can be spotted after that.

The characteristic of the warping path is just another approach of frame compression and expansion describes earlier. Thus, frame fixing is a mean of solution to speech frame variations and it still preserve the global distance score; the DTW-FF algorithm only make adjustment on the feature vectors of the horizontal and vertical local distance movements, leaving the diagonal movements as it is with their respective reference vectors. This would

essentially help to reduce the patterns with which the unknown input has to be compared. The frame fixing is done throughout the samples, also taking considerations population sample which has same number of frames among them, as the averaged frames for the reference template.

As for the recognition comparison between typical DTW and DTW-FF algorithm, the results in Table 1 showed the same recognition rate due to same pattern matching between frames template of both algorithms of Malay language digits 0-9. Moreover, the global distance score is preserved, this makes a stronger argument that the recognition before and after DTW-FF is identical and no loss of information during the DTW-FF algorithm.

Table 1: Recognition between typical DTW and DTW-FF

Subject	DTW (%)	DTW-FF (%)
1	92	92
2	92	92
3	90	90
4	84	84
5	84	84

V. CONCLUSION

The time alignment based on DTW method for pre-processing LP coefficients and a pitch optimization is described in this paper. The result from experiments conducted shown that the DTW-FF algorithm can perform frame matching between an input and a reference speech frame as good as typical DTW algorithm. From this obtained result, further use of the fixed frame speech along with pitch feature can be applied to neural network speech recognition utilizing the local distance scores. In conclusion, the DTW-FF algorithm can be used to see the characteristic of the word of speaker. This is an alternative method found to resolve the problem of data feeding into neural network algorithm or other subsequent pattern matching using the well known DP method. Warping path can show the characteristics of speaker or words spoken. This information can be used together with pitch to study speaker recognition or word recognition and this is a news to improve further the performance.

REFERENCES

[1] S.H.Salleh. *An Evaluation of Preprocessors for Neural Network Speaker Verification*. University of Edinburgh, UK: Ph.D. Thesis, 1997.

[2] N.M.Botros and S.Premnath. Speech Recognition using Dynamic Neural Networks. *International Joint Conference in Neural Network*. 4: 737-742, June 1992.

[3] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition, *IEEE Transactions on* ASSP-26(1): 43-49. February 1978.

[4] L.Rabiner and B.H.Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[5] M.H.Kuhn, H.Tomaschewski, and H.Ney. Fast Nonlinear Time Alignment for Isolated Word Recognition. *Proceedings of ICASSP*. 6:736-740, April 1981.

[6] M.J.Creany. *Isolated Word Recognition using Reduced Connectivity Neural Networks with Non-Linear Time Alignment Methods*. University of New Castle-Upon-Tyne: Ph.D. Thesis, 1996.

[7] S.Sae-Tang and C.Tanprasert. Feature Windowing for Thai Text-Dependent Speaker Identification using MLP with Back-Propagation Algorithm. *IEEE International Symposium on Circuits and Systems*, Geneva. 3: 579-582, May 2000.

[8] W.H.Abdulla, D.Chow and G.Sin. Cross-Words Reference Template for DTW-Based Speech Recognition System. *TENCON*. Bangalore, India, 1: 1-4, 2003.

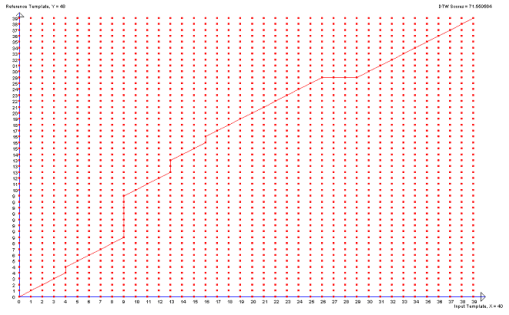
[9] Y.Matsuura, H.Miyazawa, and T.E.Skinner. Word Recognition using a Neural Network and a Phonetically Based DTW. *Proceedings of the 1994 Workshop in Neural Networks for Signal Processing*, 329-334, September 1994.

[10] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.J.Lang. Phoneme Recognition using Time Delay Neural networks. IEEE Transactions on ASSP, 37(3): 328-338, March 1989.

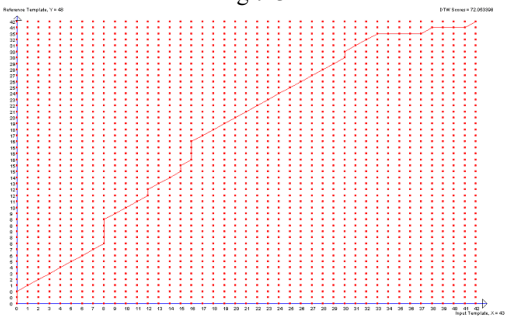
[11] S.Uma, V.Sridhar, and G.Krishna. Time-Normalization Techniques for Speaker-Independent Isolated Word Recognition. *Proceedings of Pattern Recognition Conference: Image, Speech and Signal Analysis*. 3: 537-540, Sep 1992.

[12] C.S.Myers and L.R.Rabiner (1981). Connected Digit Recognition Using a Level-Building DTW Algorithm. *IEEE Transactions on ASSP,* 29(3): 351-363.

[13] E.F.Cabral Jr. and G.D.Tattersall. Trace-Segmentation of Isolated Utterances for Speech Recognition. *International Conference on ASSP* 1:365-368, May 1995.



Digit ‘3’



Digit ‘7’

Figure 2: Sample pattern of digit ‘3’ and digit ‘7’